D3.1 First online report on available tools

Succeed

29/07/2013

Abstract
This deliverable produces a survey of existing tools, ground truth data and lexicon data for digitisation. The survey is presented as a list which will be maintained throughout the project.

Succeed is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

Document information

Deliverable number	D3.1	Start: 1	Due: 6	Actual: 7
Deliverable name	First online report on available tools			
Internal/External	External			
Activity type	SUPP			
Participant	UA, INL, IAIS, PSNC, KB			
Estimated person months per participant for this deliverable	2PM for the Deliverable. KB was originally not included. The number per participant will be higher.			
Dissemination level[1]	RP			

Document history

Revisions				
Version	**Status**	**Author**	**Date**	**Changes**
0.1	Draft	Sebastian Kirch	6-6-2013	Initial version
0.2	Draft	Sebastian Kirch	17-6-2013	Marion Borowski and Katrien Depuydt
0.3	Draft	Sebastian Kirch	21-6-2013	Feedback WP3 members
0.4	Draft	Katrien Depuydt	24-6-2013	Minor Changes
1.0	Final	Katrien Depuydt	29-7-2013	Final version consolidated

Approvals				
Version	**Date of approval**	**Name**	**Role in project**	**Signature**
0.4	26/07/2013	Aly Conteh	Internal supervisor	OK
1.0	29/07/2013	Isabel Martínez	Technical Project Manager	OK

Distribution			
This document was sent to:			
Version	**Date of sending**	**Name**	**Role in project**
0.1	6-6-2013	Marion Borowski, Katrien Depuydt	WP3 co-lead

1 PU Public; RP Restricted to other programme participants (including Commission Services); RE Restricted to a group specified by the consortium (including Commission Services); CO Confidential, only for members of the consortium (including the Commission Services)

Succeed is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

0.2	17-6-2013	Sebastian Kirch, Marion Borowski, Bob Boelhouwer, Jesse de Does, Lotte Wilms, Clemens Neudecker, Isabel Martinez, Tomasz Parkola	WP3 members
0.3	25-6-2013	Sebastian Kirch	WP3 co-lead
0.4	25-6-2013	Aly Conteh	Internal reviewer
1.0	26-7-2013	Isabel Martínez	Technical Project Manager
1.0	29-7-2013	Rafael Carrasco	Project Coordinator

Succeed is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

Table of Contents

INTRODUCTION

This deliverable is part of WP3. This work package will support the validation of digitisation tools, linguistic tools and resources created by research and development programs and their transference for exploitation in libraries and other cultural heritage organisations. In particular, the results from previous and ongoing research projects (e.g. IMPACT, CONTENTUS) will be made available to the community and the partners involved will provide assistance for the adaptation of the tools to specific domains and languages as well as training in the usage of tools.

The objective of this deliverable is to produce a survey of existing tools, ground truth data and lexicon data for digitisation. The survey is presented as a list which will be maintained throughout the project (Deliverable 3.2: Final online report on available tools) by monitoring the research outcomes in digitisation techniques. It will focus on two main groups: image processing tools and text processing tools.

To support the take-up and validation of digitisation tools and resources, some of the tools in this list will be implemented and tested at each of the four participant libraries in this project and also at external libraries. From this survey, possible candidate tools for implementation will be selected, according to well-defined criteria. This deliverable contains the complete tool overview as well as information on which tools were retained for possible further implementation.

PROCEDURE

The following procedure was implemented to collect, organise and select the tools for this deliverable:

1) Tool list: Description and Taxonomy

The first step in the selection process was to define a hierarchical taxonomy for categorising the tools. This taxonomy is based on a simplified digitisation workflow including the following steps:

- **Image Processing**
  Algorithms for OCR or layout analysis rely heavily on the quality of the input images. Scanned documents might be distorted, skewed or contain noise artifacts that prevent these algorithms from producing good results. Additionally, working with bitonal images is generally held to be the most efficient practice for document analysis.
  The purpose of this step is to enhance the quality of the scanned documents, both for visual presentation in digital libraries and eBooks and to improve the results of the subsequent steps such as OCR.

- **Text Recognition**
  Text recognition or Optical Character Recognition (OCR) is usually the core element of a digitisation workflow. It enables users to perform full-text searches on scanned documents which were not accessible to this kind of search beforehand. The OCR results are also the foundation for a lot of subsequent processing steps such as layout analysis and text processing. In addition to standard OCR processing, this step might also include a manual postcorrection of OCR results to improve their quality.

- **Layout Analysis**
  Documents such as books, newspapers or magazines are usually a composition of various structural elements such as images, tables, headings or articles. In this step of the digitisation workflow these structural elements are automatically detected and reconstructed to allow for a more fine grained search that can for example be restricted to headings or image subtitles.

- **Text Processing**
  The purpose this step is to make the digitised text more accessible to users and researchers by applying linguistic resources and language technology. For example lexical resources for retrieval and OCR can be constructed. Relevant technologies are: dealing with spelling and morphological variation for historical text, integration of linguistic resources in retrieval in library infrastructure, named entity recognition.

- **Metadata Processing**
  Metadata for digitized documents can often be derived from existing databases or catalogue systems. Additionally, metadata is generated in the previous processing steps of the digitisation workflow (text recognition, named entity recognition, ...). To be conforming to existing digitisation standards or to integrate this information into a digital library portal, metadata often has to be transformed into specific data formats which is done in this step of the workflow.

- **Evaluation**
  In order to compare or to improve OCR, layout or NLP results, these results need to be evaluated using specific tools and resources. In digital imaging and OCR, ground truth is the objective verification of the properties of a digital image, used to test the accuracy of automated image analysis processes (The ground truth of an image's text content, for instance, is the complete and accurate record of every character and word in the image).

For each of these steps there is a *group* for tools supporting the corresponding tasks in the digitisation workflow. Additionally, tools that do not fit in any of these steps are added to the *group* "Miscellaneous Utilities". These tool *groups* are the first hierarchy level in the tool taxonomy.

To allow for a finer grained categorisation, the tool *groups* are further divided into *types* and *subtypes*. These subcategories have been defined individually for each *group* based on the exact purpose of a tool. The following listing outlines the complete taxonomy including *types* (2nd hierarchy level) and *subtypes* (3rd hierarchy level):

Image processing

- **Image Processing and Enhancement**
  Binarisation, geometric correction, noise removal, ...
- **Image Segmentation**
  Region/Block/Line/Word/Character segmentation

Text Recognition

- **Core Text Recognition**
  Recognition engines: Printed, handwritten, typewritten and other like music / formulas / pictograms
  Utilities/procedures for training and customisation: Training interface of finereader/ dictionary use for OCR / etc. This will include topics like "Retraining tesseract"
- **Postcorrection**
  Automated; Semi-automated; Manual

Layout Analysis

Tables, headlines, table of contents, footnotes, newspaper articles

Text Processing

- **NLP Tools**
  Keyword Extraction, Language Identification, Lemmatization, Lemmatizer, NER, NLP toolset and resources, POS Tagger, Sentiment Mining, Spelling variation, Stemmer/Lemmatizer, Text Classification, Tokenizer

Metadata Processing

Conversions; enhancements; Linked (open) data; Format libraries like for METS; Normalisation (like date conversion);

Evaluation

OCR (text) / Layout / NLP tool evaluation

Miscellaneous Utilities

Image conversion tools / tools for creating presentation versions etc; Format conversion; compression etc

In addition to the categories of the taxonomy, each tool in the list is described by various attributes to provide more information on what the tool does and how it can be used to support the digitisation process. Not all of these attributes are mandatory for the overview. Some are only relevant for a specific type of tool (language support), others are only interesting for the evaluation at libraries (technical context, time and effort for installation). The following table gives an overview of these attributes:

Attribute	Mandatory	Description
Name of the tool	Yes	Tool name
Description	Yes	Short description of the tool: function, usage scenarios, etc. In most cases this is the description provided by the tool authors.
Link to the tool/website	Yes	Link to website with more information about the tool.
Entry author	Yes	The author (WP3 participant) who added the tool to the list.
Type of license	Yes	License under which the tool is distributed.
Language support	No	Some tools only support a restricted set of languages. This attribute is only relevant for these tools.
Technical context	No	E.g. Programming language, type (command line, web service, API, etc). This information is only provided for tools in the final list.
Time and effort for installation	No	Rough estimation on how long it takes to install the tool. This information is only provided for tools in the final list.
Name of the tool		Tool name

2) Tool list: Complete survey

After defining tool categories the search for relevant tools began. A Google Docs spreadsheet was used to collect the tools and to organise them into the categories specified by the taxonomy. The goal was to come up with a list as complete as possible for the given categories. The sources for research included mainly web resources, e.g. research project/group websites, existing tool overviews and plain web searches using search engines. The WP3 partners added their results to the spreadsheet individually.

3) Criteria for first selection

The tool list serves two purposes: One is to provide a complete overview of research and open source tools to support mass digitisation as stated in the description of this deliverable. The other is to support libraries in choosing appropriate tools for evaluation in task 3.3 (take-up support). For these libraries, additional filters are necessary to make the tool list smaller, more relevant and more manageable. The goal was to come up with 3-5 tools per category.

Based on the discussion with the candidate libraries for tool evaluation (internal and external to the project), a three-step process was defined to filter out tools that are most probably not relevant for the libraries. For each step there is a "criterion for exclusion" which can either be "yes" or "no", "yes" meaning that the tool stays in the list and "no" that the tool is discarded. These criteria were consecutively applied to the tools list to come up with a final list to be presented to the libraries for evaluation. The criteria are:

Criterion	Description
Trial version available?	Is there a free trial version available to test the tool within Succeed?
Documentation/Support available?	Is there technical documentation or support available?
Information assuring tool performance available?	Is there information about the tool being used in other projects, information about existing benchmarks or information from users about the tool available?

Finally, there are two versions of the tool list for deliverable 3.1: a filtered version for tool evaluation and an unfiltered version which is made publicly available.

4) Criteria for further selection

For those categories which still had too many tools there was a need for further selection. However, since the categories and the tools themselves are very diverse it is not feasible to define common criteria for further selection based on the attributes of the tools. Therefore we relied on the expertise of the WP3 members to establish a rating for each tool. This rating indicates how relevant a specific tool might be for the libraries evaluating the tools on a scale from 0 to 5 with 5 being very relevant and 0 being not relevant at all. For each tool in the categories that had to be filtered (those with more than 3-5 tools) the rating was derived in a conference call involving at least 3 of the WP3 experts in the corresponding field. In addition to the rating, a short description is provided explaining how the rating was derived. Only the tools with the best rating were selected for the final shortlist. A slightly different approach was taken for the NLP tools. The original list of NLP tools was substantial. Apart from relying on the expertise of the WP3 members, the following criteria for further selection have been taken into consideration:

- Robustness: how well can the tools deal with noisy data: e.g. syntactic parsers have been excluded for that reason.
- Number of languages the tool can process.
- Is the tool trainable for other languages?
- Maturity: research prototypes were not included in the final selection.

As for the ranking of the NLP tools: no internal ranking between the selected tools was applied. A tool was either selected (ranked 4) or out (ranked 0). A further selection of the tools in a specific category will be made according to the requirements of the libraries.

SHORTLIST

Image processing

- **Image Processing and Enhancement**

Tool Name/Link	Description	Type of license
ImageMagick / GraphicsMagick	ImageMagick is a software suite to create, edit, compose, or convert bitmap images. GraphicsMagick is the swiss army knife of image processing. It has been derived from ImageMagick 5.5.2	Apache License v2 / MIT
GIMP	GIMP is the GNU Image Manipulation Program. It is a freely distributed piece of software for such tasks as photo retouching, image composition and image authoring.	GPL
Scan Tailor	Scan Tailor is an interactive post-processing tool for scanned pages. It performs operations such as page splitting, deskewing, adding/removing borders, and others.	GPL v3
Unpaper	Unpaper is a post-processing tool for scanned sheets of paper, especially for book pages that have been scanned from previously created photocopies. The main purpose is to make scanned book pages better readable on screen after conversion to PDF. Additionally, unpaper might be useful to enhance the quality of scanned pages before performing optical character recognition (OCR).	GPL
Document Deskewer	generic skew detection and correction (for the full range 0-360 degrees) for documents printed using Roman scripts	commercial

- **Image Segmentation**

Tool Name	Description	Type of license
Character Segmentation	The developed methodology takes as input isolated words and separates them into characters.	commercial
Line and Word Segmentation	Segmentation of text regions into text lines and words independent of text recognition (OCR).	commercial

Succeed is supported by the European Union under FP7-ICT and coordinated by Universidad de Alicante.

Text Recognition

- **Core Text Recognition**

Tool Name	Description	Type of license
Abbyy FineReader Engine	State-of-the-art OCR engine	commercial
Tesseract	Tesseract is probably the most accurate open source OCR engine available	Apache License v2
Gamera OCR	OCR toolkit for Gamera: This is a Gamera toolkit for building standard text recognition applications. It is based on the Gamera framework and requires a working Gamera installation.	GPLv2
OmniPage	State-of-the-art OCR engine	commercial
ReadIris	Readiris is a OCR solution designed for private users and small to large office users	commercial

- **Postcorrection**

Tool Name	Description	Type of license
Korrektor	GUI-based software for viewing and correcting document analysis results	commercial
Cutouts	Cutouts is a web application which allows to crowdsource preparation of training data for Tesseract OCR engine.	free
Virtual Transcription Laboratory	Virtual Transcription Laboratory is Virtual Research Environment which works as a crowdsourcing platform for developing high quality textual representations of digital documents. It gives access to online OCR service and easy to use transcription editor. Images can be imported from various sources including direct import from digital libraries.	free
ALTO-Edit	ALTO Editor for text and segmentation	GPL

Layout Analysis

Tool Name	Description	Type of license
Fraunhofer Newspaper Segmenter	Award-winning (e.g. ICDAR'09,'11) page and article segmentation for scanned documents featuring complex layouts (e.g. (historical) newspapers, contemporary magazines, text books, etc.)	commercial
Functional Extension Parser	The Functional Extension Parser (FEP) is a Document Understanding Software tool capable of decoding layout elements of books. Based on the output of Optical Character Recognition, layout elements such as page numbers, running titles, headings, and footnotes are detected and	SLA

	annotated.	
Olena	A platform dedicated to image processing and pattern recognition. Its core component is a generic and efficient C++ library called Milena. Milena provides a framework to implement simple, fast, safe, reusable and extensible image processing tool chains.	GPLv2

Text Processing

- **NLP Tools**

Tool Name	Subtype	Description	Type of license
Alchemy API	Keyword Extraction	AlchemyAPI is capable of extracting topic keywords from your HTML, text, or web-based content. We employ sophisticated statistical algorithms and natural language processing technology to analyze your data, extracting keywords that can be utilized to index content, generate tag clouds, and more!	commercial
FreeLing	Language Identification	It compares the given text with available models for different languages, and returns the most likely language the text is written in. It can be used as a preprocess to determine which data files are to be used to analyze the text.	GPL
LingPipe	Language Identification	LingPipe's text classifiers learn by example. For each language being classified, a sample of text is used as training data. LingPipe learns the distribution of characters per language using character language models. Character language models provide state-of-the-art accuracy for text classification. Character-level models are particularly well-suited to language ID because they do not require tokenized input; tokenizers are often language-specific.	free
Rosette	Language Identification	Automatically Detects the Language of Any Digital Text. Rosette® Language Identifier analyzes text, identifying the language and the character encoding scheme. Detecting the language of documents is a critical first step in any process that handles multilingual text. Our software recognizes 55 languages and 45 encodings and processes files extremely quickly and accurately.	commercial
Rosette Linguistic Platform	Language Identification	Rosette® Language Identifier analyzes text, identifying the language and the character encoding scheme. Detecting the language of documents is a critical first step in any process that handles multilingual text. Our software recognizes 55 languages and 45 encodings and processes files extremely quickly and accurately.	commercial
Xerox	Language Identification	This service will tell you the language your document is written in. Language identification is often the first, necessary step in a whole line of document	commercial

		processing.	
Impact Tools	Lemmatization	IMPACT provides tools for: 1. Reducing historical word forms to one or several possible modern lemma's (lemmatization) 2. Expanding lemma lists with part of speech information to possible ("hypothetical") full forms.	ASL 2.0
Rosette Base Linguistics	Lemmatization	Sophisticated morphological analysis, segmentation, and tagging of Arabic, Asian, and European language text	commercial
FreeLing	Lemmatizer	This module is somehow different of the other modules, since it doesn't enrich the given text. It compares the given text with available models for different languages, and returns the most likely language the text is written in. It can be used as a preprocess to determine which data files are to be used to analyze the text.	GPL
Corpus Based Lexicon Tool (CoBaLT)	Lexicon building	Corpus Based Lexicon Tool (CoBaLT). A tool for corpus-based lexicon construction. Users can upload a text dataset (corpus) for use in creating an attestation-based lexicon. This tool is used to manually correct the automatically lemmatized corpus text. Verified lemmatized words plus the context in which they appear will be stored in the Information Retrieval Lexicon. The tool can handle plain text and various XML formats, among which the IMPACT Page XML format and TEI. An important requirement of the tool is that it should be fit to quickly process large quantities of data, that it is a web application that can be run from any computer in the local network, that frequent input actions can be performed with the keyboard, and that the information is presented in such a way that quick evaluation is possible.	ASL 2.0
DBPedia spotlight	NE linking	DBpedia Spotlight is a tool for automatically annotating mentions of DBpedia resources in text, providing a solution for linking unstructured information sources to the Linked Open Data cloud through DBpedia. DBpedia Spotlight recognizes that names of concepts or entities have been mentioned (e.g. "Michael Jordan"), and subsequently matches these names to unique identifiers (e.g. dbpedia:Michael_I._Jordan, the machine learning professor or dbpedia:Michael_Jordan the basketball player). It can also be used for building your solution for Named Entity Recognition, Keyphrase Extraction, Tagging, etc. amongst other information extraction tasks.	free
Apache openNLP	NER	The Name Finder can detect named entities and numbers in text. To be able to detect entities the Name Finder needs a model. The model is dependent on the language and entity type it was trained for. The OpenNLP projects offers a number of pre-trained name finder models which are trained on various freely available corpora. They can be downloaded at our model download page. To find	Apache License 2

		names in raw text the text must be segmented into tokens and sentences. A detailed description is given in the sentence detector and tokenizer tutorial. Its important that the tokenization for the training data and the input text is identical.	
FreeLing	NER	There are two different modules able to perform NE recognition. They can be instantiated directly, or via a wrapper that will create the right module depending on the configuration file.	GPL
LingPipe	NER	LingPipe is tool kit for processing text using computational linguistics. LingPipe is used to do tasks like: Find the names of people, organizations or locations in news, Automatically classify Twitter search results into categories, Suggest correct spellings of queries	Limited version free, production version at a fee
NERT	NER	NERT is a tool that can mark and extract named entities (persons, locations and organizations) from a text file. It uses a supervised learning technique, which means it has to be trained with a manually tagged training file before it is applied to other text. In addition, version 2.0 of the tool and higher also comes with a named entity matcher module, with which it is possible to group variants or to assign modern word forms of named entities to old spelling variants. As a basis for the tool in this package, the named entity re cognizer from Stanford University is used. This tool has been extended for use in IMPACT. Among the extensions is the aforementioned matcher module, and a module that reduces spelling variation within the used data, thus leading to improved performance.	GPLv2
NLTK	NER		free
Rosette Entity Extractor (REX)	NER	Identify Names, Places, Organizations, and Other Entities in Your Text	commercial
Stanford NER	NER	Stanford NER (also known as CRFClassifier) is a Java implementation of a Named Entity Recognizer. Named Entity Recognition (NER) labels sequences of words in a text which are the names of things, such as person and company names, or gene and protein names. The software provides a general (arbitrary order) implementation of linear chain Conditional Random Field (CRF) sequence models, coupled with well-engineered feature extractors for Named Entity Recognition. (CRF models were pioneered by Lafferty, McCallum, and Pereira (2001); see Sutton and McCallum (2006) for a better introduction.) Included with the download are good 3 class (PERSON, ORGANIZATION, LOCATION) named entity recognizers for English (in versions with and without additional	free

		distributional similarity features) and another pair of models trained on the CoNLL 2003 English training data. The distributional similarity features improve performance but the models require considerably more memory.	
Alchemy API	NLP toolset and resources	AlchemyAPI uses natural language processing technology and machine learning algorithms to extract semantic meta-data from content, such as information on people, places, companies, topics, facts, relationships, authors, and languages.	commercial
FreeLing	NLP toolset and resources	FreeLing is a library providing language analysis services, oriented to satisfy the needs of Natural Language Processing. FreeLing is designed to be used as an external library from any application requiring this kind of services. Nevertheless, a simple main program is also provided as a basic interface to the library, which enables the user to analyze text files from the command line. Actually, many users do not develop on FreeLing, but use it as a text processing tool.	GPL
LingPipe	NLP toolset and resources	LingPipe is tool kit for processing text using computational linguistics.	free/ commercial
Rosette Linguistic Platform	NLP toolset and resources	Comprehensive linguistic analysis of unstructured text in Asian, European and Middle Eastern languages for enhancing information retrieval, text mining, and other applications.	commercial
Apache openNLP	POS Tagger	The Part of Speech Tagger marks tokens with their corresponding word type based on the token itself and the context of the token. A token might have multiple pos tags depending on the token and the context. The OpenNLP POS Tagger uses a probability model to predict the correct pos tag out of the tag set. To limit the possible tags for a token a tag dictionary can be used which increases the tagging and runtime performance of the tagger.	Apache License 2
FreeLing	POS Tagger	There are two different modules able to perform PoS tagging. The application should decide which method is to be used, and instantiate the right class. The first PoS tagger is the hmm_tagger class, which is a classical trigam Markovian tagger, following [#!brants00!#].The second module, named relax_tagger, is a hybrid system capable to integrate statistical and hand-coded knowledge, following [#!padro98a!#].	GPL
FreeLing	POS Tagger	There are two different modules able to perform PoS tagging. The application should decide which method is to be used, and instantiate the right class.The first PoS tagger is the hmm_tagger class, which is a classical trigam Markovian tagger, following [#!brants00!#].The second module, named relax_tagger, is a hybrid system	GPL

		capable to integrate statistical and hand-coded knowledge, following [#!padro98a!#].	
NLTK Taggers	POS Tagger	This package defines several taggers, which take a token list (typically a sentence), assign a tag to each token, and return the resulting list of tagged tokens. Most of the taggers are built automatically based on a training corpus.	free, open source
Rosette Base Linguistics	POS tagger	Sophisticated morphological analysis, segmentation, and tagging of Arabic, Asian, and European language text	commercial
Stanford Log-linear Part-Of-Speech Tagger	POS tagger	A Part-Of-Speech Tagger (POS Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc., although generally computational applications use more fine-grained POS tags like 'noun-plural'. This software is a Java implementation of the log-linear part-of-speech taggers described in these papers (if citing just one paper, cite the 2003 one):	GPL2
Alchemy API	Sentiment Mining	AlchemyAPI provides easy-to-use mechanisms to identify positive / negative sentiment within any document or web page. AlchemyAPI Sentiment Analysis APIs are capable of computing document-level sentiment, user-targeted sentiment, entity-level sentiment, and keyword-level sentiment. Multiple modes of sentiment analysis provide for a variety of use cases ranging from social media monitoring to trend analysis.	commercial
Impact Tools	Spelling variations	The spelling of words in historical texts can differ widely from modern spelling. There are two general approaches to match different spellings. First, it is possible to use rewrite rules that transform words in one spelling to another. For historical dictionary which covers a large timespan, and in which variation is not limited to orthography, this approach is not satisfactory. Therefore, the use of statistics is often needed.	ASL 2.0
NLTK Stemmers	Stemmer/Lemmatizer	Interfaces used to remove morphological affixes from words, leaving only the word stem. Stemming algorithms aim to remove those affixes required for eg. grammatical role, tense, derivational morphology leaving only the stem of the word. This is a difficult problem due to irregular words (eg. common verbs in English), complicated morphological rules, and part-of-speech and sense ambiguities (eg. ceil- is not the stem of ceiling).	free, open source
Alchemy API	Text Classification	AlchemyAPI is capable of categorizing your HTML, or web-based content. We employ sophisticated statistical algorithms and natural language processing technology to analyze your information, assigning the most likely topic category (news, sports, business, etc.).	commercial
Apache	Tokenizer	The OpenNLP Tokenizers segment an input character sequence into tokens.	Apache License 2

openNLP		Tokens are usually words, punctuation, numbers, etc.	
FreeLing	Tokenizer	Tokenization rules are regular expressions that are matched against the beggining of the text line being processed. The first matching rule is used to extract the token, the matching substring is deleted from the line, and the process is repeated until the line is empty.	GPL
LingPipe	Tokenizer	Part-of-speech tagging is a process whereby tokens are sequentially labeled with syntactic labels, such as "finite verb" or "gerund" or "subordinating conjunction". This tutorial shows how to train a part-of-speech tagger and compile its model to a file, how to load a compiled model from a file and perform part-of-speech tagging, and finally, how to evaluate and tune models.	unknown
NLTK	Tokenizer	Tokenizers divide strings into lists of substrings. For example, tokenizers can be used to find the list of sentences or words in a string.	free
Rosette Base Linguistics	Tokenizer	Sophisticated morphological analysis, segmentation, and tagging of Arabic, Asian, and European language text	commercial
NLTK Classify Package	Topic Modelling	Classes and interfaces for labeling tokens with category labels (or “class labels”). Typically, labels are represented with strings (such as 'health' or 'sports'). Classifiers can be used to perform a wide range of classification tasks. For example, classifiers can be used...	free, open source

Metadata Processing

Tool Name	Description	Type of license
OxGarage	OxGarage is an web, and RESTful, service to manage the transformation of documents between a variety of formats. The majority of transformations use the Text Encoding Initiative format as a pivot format	unknown
Pandoc	Format conversion engine	GNU GPL
Augmented SIP Creator (ASC)	The ASC uses XSL scripts to transform Metadata from a source to a target XML format. It can be used to normalize and validate input metadata from heterogenous sources.	commercial
jmet2ont	A tool that makes it possible to transform metadata from a traditional XML-based schema to RDF/OWL. Mappings are described with XML. Existing mappings used in SYNAT transform traditional library/museum formats to the CIDOC CRM/FRBRoo ontology.	GPL
abbott	Abbot is a tool for undertaking large-scale conversion of XML document collections in order to make them interoperable with one another. Java technology.	own

Evaluation

Tool Name	Description	Type of license
Evaluation Tool for OCR	This tool evaluates the performance of an optical character recognition system on character and word level.	unknown
Aletheia	GUI-based document layout and text ground truthing system: a comprehensive tool for semi-automated production of ground truth and annotation of document images on page level	commercial
ISRI Tools	Images and Ground Truth text and zone files for several thousand English and some Spanish pages that were used in the UNLV/ISRI annual tests of OCR accuracy between 1992 and 1996. Source code of OCR evaluation tools used in the UNLV/ISRI annual tests of OCR Accuracy.	ASL 2.0
Layout Evaluation	Performance evaluation tool for layout analysis and segmentation methods based on detailed metrics (types of errors such as merges, splits, missed regions, etc.) and use scenarios	unknown

Miscellaneous Utilities

Tool Name	Description	Type of license
DigitLab	DigitLab (http://digitlab.psnc.pl) is an especially adapted operating system based on Linux Ubuntu. The main aim of its creation was to create a complete system which can be used for collections digitisation with the usage of free and widely available tools. DigitLab is a perfect solution for both everyday work and hands-on trainings. It allows to work with images, textual content (OCR included) and audio-visual collections. Gives access to three example digital libraries based on DSpace, dLibra and Greenstone.	free
hOCR tools	hOCR is a format for representing OCR output, including layout information, character confidences, bounding boxes, and style information. It embeds this information invisibly in standard HTML. By building on standard HTML, it automatically inherits well-defined support for most scripts, languages, and common layout options. Furthermore, unlike previous OCR formats, the recognized text and OCR-related information co-exist in the same file and survives editing and manipulation. hOCR markup is independent of the presentation.	ASL 2.0
BlackLight	Blacklight is an open source Ruby on Rails gem that provides a discovery interface for any Solr index.	CC Attribution-Share Alike 3.0 United States License.
DjVu tools	Suit of open source tools and utilities related to the DjVu format	unknown
FromThePage	FromThePage is an open-source tool that allows volunteers to collaborate to transcribe handwritten documents.	AGPL
Islandora	Javascript based TEI Transcription Editor	unknown

pyBossa	Open-source crowd-sourcing (microtasking) platform with a focus on volunteer contribution and making it super-easy to create a crowd-sourcing app.	GPLv3
Scribe	Scribe is a framework for generating crowd sources transcriptions of image based documents. It provides a system for generating templates which combined with a magnification tool guide a user through the process of transcribing an asset (an image).	ASL 2.0
tb-transcription-desk	MediaWiki based environment for a distributed, collaborative transcription effort.	GPLv2
Textlab	An innovative image and text mark-up tool, TextLab is based on the protocols of fluid text editing of revision. Here, "revision sites" are any areas of interest on a manuscript leaf or print page that indicates evidence of revision.	unknown

Along the project the survey on tools will be updated on the Succeed project website at www.succeed-project.eu/publications/available-tools/index-succeed